ChinaXiv合作期刊

一种基于双向 LSTM 的联合学习的中文分词方法 *

章登义，胡　思，徐爱萍

(武汉大学 计算机学院, 武汉 430072)

摘　要：中文分词是中文自然语言处理任务的关键技术之一。针对现有的基于深度学习的神经网络模型通常都是对单一的语料库进行训练学习，提出了一种大规模的多语料库联合学习的中文分词方法。语料库分别为简体中文数据集（PKU、MSRA、CTB6）和繁体中文数据集（CITYU、AS）。每一个数据集的输入语句的句首和句尾分别添加一对标志符。应用 BLSTM（双向长短时记忆模型）和 CRF（条件随机场模型）对数据集进行单独训练和多语料库共同训练的实验，结果表明大规模的多语料库共同学习训练能取得良好的分词效果。

关键词：中文分词；大规模语料库；联合学习；双向长短时记忆模型

中图分类号：TP391　　doi: 10.3969/j.issn.1001-3695.2018.03.0239

Joint learning method based on BLSTM for Chinese word segmentation

Zhang Dengyi, Hu Si, Xu Aiping

(*School of Computer*, *Wuhan University*, *Wuhan* 430072, *China*)

Abstract: Chinese word segmentation is one of the key technologies of Chinese natural language processing tasks. The existing neural network models based on deep learning are usually trained on single criterion corpora. This paper proposed a joint learning method based on bi-directional long short-term memory neural network and Conditional Random Fields for large-scale corpora. The corpora were composed of simplified Chinese data sets (PKU, MSRA, CTB6) and traditional Chinese data sets (CITYU, MSR) . A pair of identifiers is added to the beginning and end of each input sentence of the data set. The results of the experiments show that the effective method has good effect on Chinese word segmentation for such large-scale corpora.

Key words: Chinese word segmentation; large-scale corpora; joint learning; long short-term memory neural network model

chinaXiv:201807.00065v1

0　引言

基于中文的自然语言处理任务中，由于汉语的书写习惯，汉语中的词语不像英语等有分隔符，因此中文分词是中文自然语言处理关键基础技术之一，是其他中文文本任务（如命名实体识别、词性标注、机器翻译等）的前期重要的预处理环节。分词的准确性对中文自然语言处理尤其重要。由于中文中存在一字多意、一词多意的情况，在不同的语境下存在不同的分词方式，中文分词一直是中文自然语言处理任务中的难点。

近年来，大多数方法都将中文分词作为一个序列标注问题[1]，对给定的一段文本，为句中每个字符分配一个标签，分词任务转换为一个有监督的分类问题。所谓“序列标注”，就是说对于一个一维线性输入序列,给线性序列中的每个元素打上标签集合中的某个标签。所以，其本质上是对线性序列中每个元素根据上下文内容进行分类的问题。一般情况下，对于自然语言处理任务来说，线性序列就是输入的文本，一个汉字就是线性序列的一个元素，而不同的 NLP 任务其标签集合代表的含义可能不太相同，但是相同的问题都是：如何根据汉字的上下文给汉字打上一个合适的标签。NLP 中的序列标注任务有，中文分词、词性标注、CHUNK 识别、命名实体识别、关键词抽取、语义角色标注。常用的标注方法有支持向量机（SVM）[2]，最大熵（maximum entropy, ME）模型[3]，隐马尔可夫（HMM）模型[4]，条件随机场（CRF）模型[5]。

随着深度学习方法的发展，一些神经网络模型也被成功应用于中文分词任务。Zheng 等人[6]首先提出了基于神经网络的分词模型，采用的 Collobert 等人[7]提出的方法进行中文分词和词性标注，将神经网络模型用于预训练词嵌入，Collobert 等人[7]设计了 SENNA 系统，利用神经网络解决英文序列标注问题。Zhao 等人[13]将非监督的学习方法应用于有监督的训练中进行中文分词任务。以邻接类别（accessor variety）作为非监督学习的分词标准，对未标注的语料进行非监督学习训练得到的分词结果作为特征项输入至 CRF 层对有标注的语料进行有监督的训练，训练效果比直接用 CRF++模型训练更好。Chen 等人[8]扩展了 LSTM（long short-term memory）长短时记忆神经网络

收稿日期：2018-03-16；**修回日期：**2018-06-12　**基金项目：**国家重点研发计划资助项目（2017YFC0803700）

作者简介：章登义（1965-），男，湖北荆州人，教授，主要研究方向为嵌入式、模式识别、计算机视觉、大数据、云计算（dyzhangwhu@163.com）；胡思（1995-），女，硕士，主要研究方向为自然语言处理、模式识别、大数据；徐爱萍，女，教授，主要研究方向为自然语言处理、人工智能、模式识别、大数据、云存储.

ChinaXiv合作期刊

模型用于中文分词任务，解决了传统的神经网络模型不能学习长距离依赖关系的问题，取得了较好的分词效果。Zhang 等人[9]提出了一种基于词向量的分词的神经网络模型，将卷积神经网络和 LSTM 结合，模型输入端特征向量包含字符嵌入(character embeddings)和预训练语料库学习到的词嵌入(word embeddings)。

许多研究表明，LSTM 神经网络模型在序列标注任务中能取得不错的效果。Huang 等人[16]第一次将双向 LSTM 和 CRF 结合，对序列标注任务中的基准数据集 benchmark tasks 进行训练。该模型在词性标注任务（数据集为 PTB）、CHUNK 识别任务（数据集为 CoNLL2000）和命名实体识别任务（数据集为 CoNLL2003）中均取得了较好的性能。双向 LSTM 可以对目标词同时学习上下文信息，CRF 层可以学习训练得到句子级的标签信息。BLSTM-CRF 模型具有较好的鲁棒性，该模型对词嵌入的依赖性更小。Lample 等人[17]运用双向 LSTM 和 CRF 模型做命名实体识别的任务，使用基于字符的词向量和基于词的词向量学习特征，获取被标记词的拼写组成和被标记词在语料库中的出现位置信息，该模型在英语、德语、荷兰语和西班牙语的语料库中的准确率达到 state-of-the-art 的效果。Cai 等人[18]应用 LSTM 模型的变体于中文分词任务，提出的模型结合了门循环（GRU）神经网络模型和 LSTM 模型，GRU 模型直接生成基于字符的词向量的训练所得到的候选的分词结果，LSTM 模型用于对获得的分词结果进行评估。结合两种模型的处理的分词任务能达到不错的效果。

尽管基于神经网络模型的方法取得了巨大的成功，但一些问题仍然没有得到很好的解决。一个显著的缺点是这些方法很少考虑到知识的整合。测试中出现 OOV（out-of-vocabulary）情况，不在训练集词表中的词在测试时不能够识别并进行分词。通常情况下，模型都是分别对不同语料库进行单一的训练学习，而没有整合语料库利用多方面的信息。事实上，由于不同的语料库的分词标准不同，也不容易将语料库整合后进行联合训练。

目前运用 BLSTM-CRF 模型的序列标注问题（POS 词性标注任务、CHUNK 识别任务、NER 命名实体识别任务）均取得了不错的效果。本文在基于 BLSTM-CRF 模型的基础上，提出了一种对多钟语料库联合训练的方法来进行中文分词任务。使用双向长短时记忆模型（BLSTM）和条件随机场模型（CRF），将中文分词任务转换为一个字符级的序列标注问题。本文中使用 SIGHAN Bakeoff 2005 的数据集 PKU、MSR、AS、CITYU 和 CTB6(Chinese TreeBank6.0 数据集)进行实验，实验 1 是分别对 5 个数据集单独学习训练，实验 2 是使用全部数据集联合学习共同训练。实验中在输入端为不同语料库的输入语句的句首和句尾各自添加一对标志符。PKU 数据集的句子标志符<PKU></PKU>；MSR 数据集的句子标志符<MSR> </MSR>；AS 数据集的句子标志符<AS> </AS>；CITYU 数据集的句子标志符<CITYU> </CITYU>；CTB6.0 数据集的句子标志符<CTB> </CTB>。AS 数据集和 CITYU 数据集是繁体中文数据集，在联合训练前，使用 HanLP 工具将繁体中文转换为简体中文。

1　中文分词神经网络模型架构

中文分词任务通常被认为是基于字符的序列标注任务。标记输入句子中的每个字符，使用标签集 T= {B,M,E,S}进行标记。B 表示一个词的开始字，M 表示一个词的中间字，E 表示一个词的结束字，S 表示单字词。

基于神经网络的序列标注任务通常由三部分组成：a）字符嵌入层，文本向量化表示；b）多个神经网络转换层；c）标签推理层。整体结构如图 1 所示。

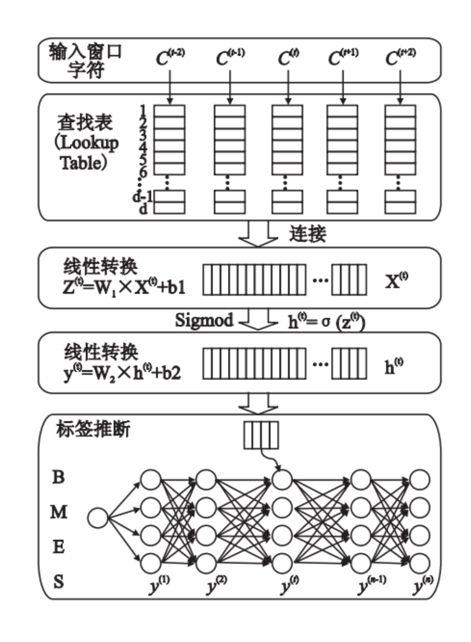


图 1　中文分词模型框架

给定长度为 n 的文本序列 $c^{(1:n)}$，大小为 k 的窗口从文本序列的第一个字 $c^{(1)}$滑动至最后一个字 $c^{(n)}$。如图 1 所示，对序列中每一个字 $c^{(t)}$，当窗口大小为 5 时，上下文信息($c^{(t-2)}$, $c^{(t-1)}$, $c^{(t)}$, $c^{(t+1)}$, $c^{(t+2)}$)输入到查询表（Lookup Table），当字的范围超过序列边界时，用两种特殊字符“<S>”和“</S>”隔开，以此保证输入字符大小固定为 k。将查询表中获得的字符向量连接成整体向量 $X^{(t)} \in R^{H_1}$，其中 $H_1 = k \times d$。$X^{(t)}$ 经过一层线性转换得到 $Z^{(t)}$,线性转换式如式（1）所示。

$$Z^{(t)} = W_1 \times X^{(t)} + b_1 \tag{1}$$

$W_1 \in R^{H_2 \times H_1}$ 是变换矩阵。H_2 是隐藏节点数，参数 $b_1 \in R^{H_2}$。激活函数 σ，通常使用 sigmoid 函数和 tanh 函数，如式（2）所示。

$$h^{(t)} = \sigma(Z^{(t)}) \tag{2}$$

再次进行线性变化，如式（3）所示。

$$y^{(t)} = W_2 \times h^{(t)} + b_2 \tag{3}$$

其中：$W_2 \in R^{D \times H_2}$，D 为词位标签词。$y^{(t)} \in R^D$，$y^{(t)}$ 中每一个元素代表对应词位标签的得分。通过对字符序列中的每个字符进行以上的计算,可以得到该字符序列中每个字符的标签得分

chinaXiv:201807.00065v1

ChinaXiv合作期刊
chinaXiv:201807.00065v1

矩阵。由于一个字符序列中,字符标签之间存在强依赖关系,因此,可以引入一个矩阵 A 来表示字符标签之间的转换关系。 A_{ij} 表示标签 i 转移到标签 j 的概率。通过后向传播算法,从训练集学习得到概率矩阵 A。

上述神经网络模型在中文分词任务中表现出较好的效果，但是由于窗口大小的限制,模型不能学习窗口外的上下文信息,丢失了长距离依赖信息，对分词的准确性有影响。而 LSTM 则突破了窗口的限制，可以利用长距离的上下文信息。

2　基于 BLSTM 的模型架构

2.1　文本向量化

利用神经网络模型处理中文分词问题，首先需要将文本向量化表示。使用一个特定维度的特征向量代表字符。字符向量可以刻画字与字在语义和语法上的相关性，并且作为字符特征成为神经网络的输入。使用训练语料集中的所有字建立一个大小为 d×N 的汉字字典矩阵，d 是字符向量的维度，N 是字典的大小，构造了一个字符到字嵌入的查找表，将输入的中文字符转换为字嵌入向量，作为模型的输入。研究表明,使用大规模无监督学习得到的字向量作为输入矩阵的初始值比随机初始化有着更优的效果[10]。本文中使用 word2vec[11] 在中文维基百科语料库预训练获得字符向量。

2.2　循环神经网络（RNN）

在传统的神经网络模型中，是从输入层到隐含层再到输出层，层与层之间是全连接的，每层之间的节点是无连接的。处理每个时刻的信息时是独立的。而循环神经网络 RNN 在隐藏层中增加节点中的互连，隐藏层的输入不仅包括输入层的输出还包括上一时刻隐藏层的输出。网络模型会对前面的信息进行记忆并应用于处理当前输出数据的计算中。RNN 的模型结构如图 2 所示。RNN 按时间顺序展开的示意图如图 3 所示。

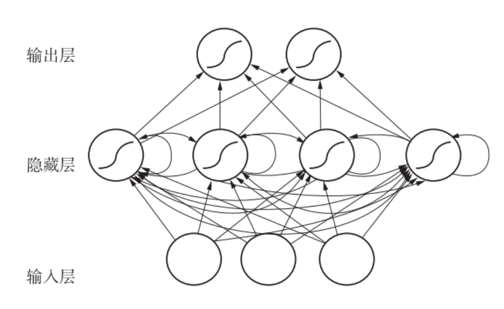


图 2　RNN 网络结构示意图

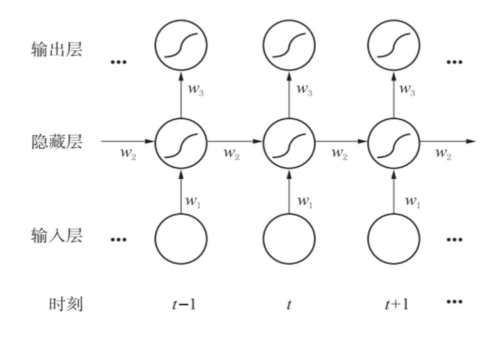


图 3　RNN 展开结构示意图

图中的每个节点表示在每个时刻 RNN 网络的一层。w_1 是输入层到隐藏层的连接权值，w_2 是上一时刻隐藏层到当前隐藏层的连接权值，w_3 是隐藏层到输出层的连接权值。RNN 中每时刻的权值都是共享的,当前时刻的输出依赖于上一时刻的输出。t 时刻隐藏层输出为

$$h^{(t)} = g(Uh^{(t-1)} + Wx^{(t)} + b) \tag{4}$$

其中：$x^{(t)}$ 是 t 时刻的输入，$h^{(t-1)}$ 是 t-1 时刻的隐藏状态，U 是输入层到隐藏层的权值矩阵,W 是隐藏层到输出层的权值矩阵，b 是偏置参数，g 是激活函数通常是 tanh 函数。

传统的 RNN 只能利用上文消息，而在许多自然语言处理任务中，需要利用上下文信息，因此扩展了双向 RNN 能够同时利用序列中的历史和未来信息。将序列信息分两个方向输入至模型中，使用两个隐藏层保存两个方向的输入信息，将隐藏层相应的输出连接到相同的输出层。BRNN 的网络结构展开示意图如图 4 所示。

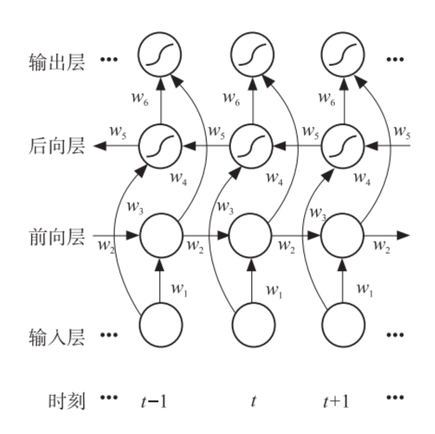


图 4　BRNN 展开结构示意图

2.3　LSTM 长短时记忆模型

长短时记忆模型（LSTM）[12]是循环神经网络的变体。尽管在理论上，RNN 可以处理任何长距离依赖问题，但实际上，由于梯度消失/爆炸问题而很难实现。LSTM 通过引入门机制和记忆单元为此提供了解决方案，用 LSTM 单元代替 RNN 中的隐藏层。LSTM 单元结构图如图 5 所示。

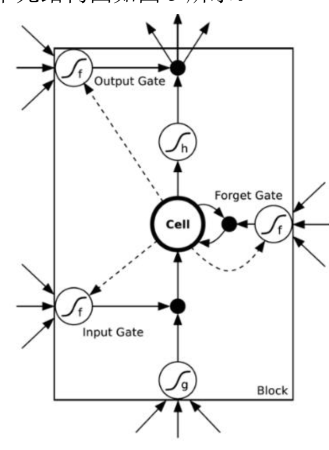


图 5　LSTM 结构图

LSTM 中保存的历史信息受输入门、遗忘门和输出门控制。x 是输入数据，h 是 LSTM 的单元输出，c 是 LSTM 记忆单元的值。输入门 **i**，遗忘 **f**，记忆单元 **c** 和输出门 **o**。

$$\tilde{c}^{(t)} = \tanh(W_{xc}x^{(t-1)} + W_{hc}h^{(t-1)} + b_c) \tag{5}$$

其中：$\tilde{c}^{(t)}$ 是当前时刻记忆单元的候选值，$x^{(t)}$ 是 t 时刻的输入数据，W_{xc} 是 t 时刻输入数据的权值，$h^{(t-1)}$ 是上一时刻 LSTM 的输出，W_{hc} 对应上一时刻 LSTM 单元的输出的权值，b_c 是偏置参数。

$$i^{(t)} = \sigma(W_{xi}x^{(t-1)} + W_{hi}h^{(t-1)} + W_{ci}c^{(t-1)} + b_i) \tag{6}$$

其中：$i^{(t)}$ 是输入门当前的状态值，$i^{(t)}$ 受当前输入数据 $x^{(t)}$、上一时刻 LSTM 的输出 $h^{(t-1)}$ 和上一时刻记忆单元值 $c^{(t-1)}$ 的影响。W_{xi}、W_{hi}、W_{ci} 分别为对应的权值。

$$f^{(t)} = \sigma(W_{xf}x^{(t-1)} + W_{hf}h^{(t-1)} + W_{cf}c^{(t-1)} + b_f) \tag{7}$$

其中：$f^{(t)}$ 是遗忘门当前的状态值，遗忘门是控制历史信息对当前记忆单元的影响。W_{xf}、W_{hf}、W_{cf} 分别是对应遗忘门的权值。

$$c^{(t)} = f^{(t)} \odot c^{(t-1)} + i^{(t)} \odot \tilde{c}^{(t)} \tag{8}$$

其中：$c^{(t)}$ 是 t 时刻记忆单元的状态值，$\odot$ 表示元素间的点积，逐点相乘。记忆单元的状态值由输入门和遗忘门共同调节。

$$o^{(t)} = \sigma(W_{xo}x^{(t-1)} + W_{ho}h^{(t-1)} + W_{co}c^{(t-1)} + b_o) \tag{9}$$

输出门的输出状态值 $o^{(t)}$，控制记忆单元状态值的输出。

$$h^{(t)} = o^{(t)} \odot \tanh(c^{(t)}) \tag{10}$$

其中：$h^{(t)}$ 是 t 时刻 LSTM 单元的输出状态值，当前时刻的隐藏状态。

LSTM 的门机制使得模型可以捕捉长距离历史信息，为了同时获取上下文信息，采用双向 LSTM，因此，BLSTM 中的隐藏状态 $h^{(t)}$ 可表示如下：

$$h^{(t)} = \overrightarrow{h^{(t)}} \oplus \overleftarrow{h^{(t)}} \tag{11}$$

其中：$\overrightarrow{h^{(t)}}$ 和 $\overleftarrow{h^{(t)}}$ 分别是 LSTM 中在 t 时刻输入数据的前向和后向的隐藏状态，$\oplus$ 表示整合操作。

2.4 标签推断 CRF 层

对于基于字符的中文分词任务，本文需要考虑相邻标签间的依赖关系。例如，B（开始）标签后面应该跟一个 M（中间）标签或者 E（结束）标签，而一个 M 标签后面不能跟一个 B 标签或者 S 标签。因此，不是独立地使用 $h^{(t)}$ 来作标签决策，而是使用条件随机场（CRF）来共同建模标签序列。

对于给定的句子 $x = (x_1, x_2, ..., x_n)$，和对应的预测的标签 $y = (y_1, y_2, ..., y_n)$，预测评估分数定义如下：

$$s(x, y) = \sum_{i=1}^{n}(A_{y_{i-1}y_i} + P_{i,y_i}) \tag{12}$$

A 是一个转换分数矩阵，$A_{i,j}$ 是衡量标签 i 到 j 的分数，P_{i,y_i} 代表字符 x_i 的第 y_i 个标签的分数。P_i 定义如下：

$$P_i = W_s h^{(t)} + b_s \tag{13}$$

其中：$h^{(t)}$ 是 BLSTM 中 t 时刻输入数据 $x^{(t)}$ 的隐藏状态，W_s 是权值矩阵，b_s 是偏置。

在 CRF 层，句子 **x** 被标记为序列 **y** 的可能性概率计算如下：

$$p(y \mid x) = \frac{e^{s(x,y)}}{\sum_{y \in Y_x} e^{s(x,y)}} \tag{14}$$

其中：Y_x 表示给定句子 x 的所有可能序列 y 的集合。训练时，本文使用最大似然估计法最大化序列 y 的 $\log(y \mid x)$ 值。解码时，预测结果输出得分最高的序列 y^*，计算公式如下：

$$y^* = \arg\max_{y \in Y_x} s(x, y) \tag{15}$$

可以使用 Viterbi 算法（一种动态规划算法）来解决训练和解码过程中的效率问题。

3 实验结果与分析

3.1 实验数据

本文使用的实验数据集是 PKU、MSR、AS、CityU（来自 SIGHAN2005）和 CTB6（来自 Chinese TreeBank 6.0）。

几个数据集的分词标准各有不同，PKU 是由北大计算语言学研究所提供的语料库。其分词特点之一是姓名中姓和名要分开，组织机构等在语法词典中的直接标记，大多数短语性的词语先切分再组合。例如 欧阳/修、中国/计算机/学会、联合国/教科文/组织 。

MSR 是微软亚洲研究院所提供的语料库，其分词特点是由大量的命名实体构成的长单词，例如 联合国教科文组织、全国人民代表大会、水利部长江水利委员会、葛洲坝集团公司。

AS 由台湾中央研究院提供的语料库，分词规范与北大制定的分词规范类似同时也与台湾地区的语言使用习惯相关，分词例子：平漢/ 鐵路 、二二八/事變 、 台北市/第一/信用/合作社。

CityU 由香港城市大学提供的语料库，分词规范受香港地区的使用习惯影响。分词例句： 本/趟/列車/共/有/八/卡/掛車。

CTB 6.0 是宾州大学汉语树库中的语料库，该语料库是经过句法标注的数据，按照句子内部结构形成的句子树。

不同的分词标准对比如表 1 所示

表 1 不同语料库分词标准对比

数据集	句子示例
PKU	王 / 明 / 到达 / 奔驰 / 公司
MSR	王明 / 到达 / 奔驰公司
AS	王明 / 到達 / 賓士 / 公司
CityU	王明 / 到達 / 平治 / 公司
CTB	王明 / 到达 / 奔驰 / 公司

实验时随机选择训练数据中的 90%的数据作为训练集，剩下 10%作为开发集。所有的数据在输入前需要经过预处理，将中文习语替换成*,英文单词替换成&，数字替换成$，在大规模的无标注的语料上进行字向量训练，将训练完成的字向量作为

chinaXiv:201807.00065v1

ChinaXiv合作期刊

chinaXiv:201807.00065v1

本次实验的词向量。将每个数据集的输入语句都添加各自的标志符，不同数据集带有标志符的句子形式示例如表 2 所示。当计算最终输出的分值时不计算标志符。为了便于评估，本文使用标准 bake-off 打分程序来计算准确率 P,召回率 R，F1 分值。

表 2　数据集输入句子形式

数据集	**输入句子示例**
PKU	<PKU>王明到达奔驰公司</PKU>
MSR	<MSR>王明到达奔驰公司</MSR>
AS	<AS>王明到噠賓士公司</AS>
CITYU	<CITYU>王明到噠平治公司</CITYU>
CTB	<CTB>王明到达奔驰公司</CTB>

实验在内存为 8 BG 的 Ubuntu 系统上进行，程序采用 Python 语言进行编程

模型中的超参数设置如表 3 所示

表 3　模型超参数设置

参数	**参数值**
上下文窗口长度	k = 5
字符向量长度	d = 100
隐藏层单元数	h = 128
初始学习率	α = 0.1
Dropout 比率	p = 0.2
Batch	b = 128

3.2　实验结果分析

实验 1　将每个数据集的训练数据分别输入至模型中，对五个数据集单独进行训练。实验结果如表 4 所示。表 4 列出了不同数据集训练的性能。

表 4　不同数据集在 BLSTM 模型上的性能　　/%

性能（%）	数据集				
	PKU	MSRA	CITYU	AS	CTB
P	95.1	96.5	95.4	95.2	95.6
R	94.6	96.3	94.8	95.3	95.1
F	95.3	96.8	95.3	94.9	96.2

表 5 中列出了本文提出的模型与其他模型的性能对比。与文献[6]、的模型、文献[8]的扩展的 LSTM 模型、文献[9]的 CNN 和 LSTM 结合的模型、文献[13]的将非监督学习方法应用于 CRF 的模型和文献[14]的快速高效的基于 BLSTM 的模型变体的实验结果 F 值对比。

表 5　不同模型在不同数据测试集上 F 值对比结果(%)

模型（%）		数据集				
		PKU	MSRA	CITYU	AS	CTB6
Zheng(2013)[6]	F	92.4	93.3	-	-	-
Chen(2015)[8]	F	95.7	96.4	-	-	94.9
Zhang[9]（2016）	F	95.7	97.7	-	-	96.0
ZHAO(2008)[13]	F	95.4	97.6	96.1	95.7	94.3
cai(2017)[14]	F	95.4	97.0	95.4	95.2	-
BLSTm	F	95.1	96.8	95.3	94.9	96.2

表 4 列出了 BLSTM 模型在不同数据集上的性能，可以看出 BLSTM 模型能达到较好的效果。MSR 数据集的各项训练性能最好。MSR 测试集的准确率高达 96.5%，PKU 的准确率达到 95.1%，CITYU 的准确率达到 95.4%，AS 的准确率达到 95.2%，CTB6.0 的准确率达到 95.6%。由表 5 可以看出，与其他模型相比，使用 BLSTM 模型对单个数据集训练可以达到较好的效果，但不是最佳的效果。Zhang[9]等人将卷积神经网络和 LSTM 结合的模型在简体中文数据集上的训练效果最好，PKU 的 F1 值高达 95.7%，MSR 的 F1 值高达 97.7%。Zhao[13]等人提出的传统的非监督学习和监督学习的方法相结合在繁体中文数据集上的效果较其他方法更好，CITYU 的 F1 值达 96.1%，AS 的 F1 值达 95.7%。而本文应用的 BLSTM 模型在单个数据集上的训练效果虽然没有达到最高的性能，但也取得了较好的分词效果，其中 CTB 的 F1 值高达 96.2%，比其他方法的 F1 值稍高。PKU 的 F1 值达到 95.1%，MSR 的 F1 值达到 96.8%，CITYU 的 F1 值达到 95.3%，AS 的 F1 值达到 94.9%，单个数据集都取得了较好的分词效果。

实验 2　对五个数据集联合训练。将带有标志符的输入语句共同输入至一个模型中进行联合学习训练。联合训练的结果与单个数据集训练的结果对比如表 6 所示。

实验 2 训练的结果与文献[15]提出的使用多个数据集进行对抗性学习方法对比，结果如表 7 所示。

表 6　单个数据集训练与多个数据集联合训练的结果对比(%)

性能（%）		数据集				
		PKU	**MSRA**	**CITYU**	**AS**	**CTB**
P	单个训练	95.1	96.8	95.4	95.2	95.6
	联合训练	**95.6**	**97.4**	**95.8**	**96.0**	**95.8**
R	单个训练	94.6	96.3	94.8	**95.3**	95.1
	联合训练	**96.2**	**97.3**	**95.2**	95.2	**96.1**
F	单个训练	95.3	96.5	95.3	**94.9**	**96.2**
	联合训练	**95.8**	**97.1**	**95.2**	94.7	95.8

表 7　多语料库联合训练实验结果对比(%)

模型（%）		数据集				
		PKU	**MSRA**	**CITYU**	**AS**	**CTB6**
CHEN(2017)[15]	P	94.9	95.9	95.4	94.2	96.0
	R	93.8	96.1	95.7	95.1	96.3
	F	94.3	96.0	**95.6**	94.6	**96.2**
联合训练	P	95.6	97.4	95.8	96.0	95.8
	R	96.2	97.3	95.4	95.2	96.1
	F	**95.8**	**97.1**	95.2	**94.7**	95.8

由表 6 可知，大规模数据集联合训练所得的结果比数据集分别在模型上训练的效果更好。联合训练时简体中文数据集 PKU、MSR 的各项性能较单独训练时均有提高，AS 的性能略有下降。联合训练 PKU 数据集的准确率 P 达 95.6%，召回率 R 值 96.2%，F1 值 95.8% ；MSRA 训练集的准确率 P 达 97.4%，

chinaXiv:201807.00065v1

召回率 R 值 97.3%，F1 值 97.1% ；CITYU 训练集的准确率 P 达 95.8%，召回率 R 值 95.2%，F1 值 95.2% ；AS 训练集的准确率 P 达 96.0%，召回率 R 值 95.2%，F1 值 94.7% ；CTB 训练集的准确率 P 达 95.8%，召回率 R 值 95.1%，F1 值 95.8%。联合训练中，PKU、MSRA 和 CITYU 的测试集的结果均较单独训练结果好。可知，联合训练中大部分数据集的训练效果较好。

由表 6 可以看出，AS 和 CTB 数据集单独训练的效果较联合训练更好，原因是 AS 语料库是台湾中央研究院提供中文繁体语料库，联合几种语料库共同训练时，本文使用数据前用 HanLP 工具将繁体中文转为简体中文，在转换过程中，可能出现词语的简繁体的不对应情况，导致联合训练时，训练结果较语料库单独进行有监督的训练的情况稍差。宾州大学汉语数库 CTB 是带标签的树库文件，语料库中的语料是经过句法标注的，是基于短语结构的 LDC 中文树库采用句子的结构成分描述句子的结构，与其他语料库的标注方法不同，是导致联合训练时，该语料的训练结果比该语料单独训练的效果差的原因。

由表 7 可知，本文中将五个数据集进行联合训练时，与 chen(2017)[15]所提出的对八种数据集进行对抗性学习训练所取得的训练结果比较，发现本文中所提出的方法与之性能相当，PKU 数据集、MSRA 数据集和 AS 数据集的训练结果略优于 chen 等人的实验结果。本文提出的方法相比 chen 等人设计的模型要更为简便，却能达到相当的结果。Chen 等人提出的模型对多种语料库对抗性学习训练的效果略逊于与单独对每个语料库进行训练的效果。对多语料库进行联合训练的研究还有很大的发展空间。

4　结束语

本文针对自然语言处理中的中文分词任务，提出的基于 BLSTM 和 CRF 结合构建了一个深度神经网络模型。针对多个不同分词标准的数据集，为不同数据集的句子加入一对各自独有的标志符用来表明数据属于哪一个数据集，对数据进行共同训练，实验结果表明该分词模型能够达到 state-of-the-art 的效果。并且与现有的提出的多标准共同训练的其他复杂分词模型相比，本文所用的模型结构更为简单有效。未来可以将该方法应用于命名实体识别等任务中。BLSTM-CRF 模型应用于多种序列标注任务取得了较好的效果，本文在应用该模型的基础上，联合多种语料库进行训练，进行中文分词任务，取得了较好的效果。目前提出的方法的模型较为简单，未来可进一步进行多标准多任务的学习方法的研究。

尽管本文提出的模型在中文分词任务中取得了比较好的效果，但仍有需要改进的地方。文中对多个数据集训练时，利用 HanLP 工具将繁体中文转换为简体中文。简繁转换中的简繁分歧词可能会影响最终的分词结果，例如打印机和印表機，以太网和乙太網，总线和匯流排等。实验也表明联合训练时，繁体中文数据集的分词效果没有其他方法训练的分词效果好。简体中文与繁体中文之间存在的差异性也是联合多语料库训练存在的问题之一。联合语料库共同训练，数据集较大，训练的时间较单个语料库训练的时间更长，所需空间更大，这也是多语料库训练的需要改进的方向。

参考文献：

[1] Xue Nianwen, Converse S P. Combining classifiers for Chinese word segmentation [C]// Proc of the 1st SIGHAN Workshop Workshop on Chinese Language Processing. Stroudsburg: Association for Computational Linguistics, 2002: 57-63.

[2] Boser B E, Guyon I M, Vapnik V N. A training algorithm for optimal margin classifiers [C]// Proc of the 5th Annual Workshop on Computational Learning Theory. New York: ACM Press, 1992: 144-152.

[3] Berger A L, Pietra V J D, Pietra S A, *et al.* A maximum entropy approach to natural language processing [J]. Computational Linguistics, 1996 22 (1): 39-71.

[4] Eddy S R. Hidden Markov models [J]. Current Opinion in Structural Biology, 1996 6 (3): 361-365.

[5] Lafferty J D, McCallum A, Pereira F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]// Proc of the 18th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc. 2001: 282-289.

[6] Zheng Xiaoqing, Chen Hanyang, Xu Tianyu. Deep learning for Chinese word segmentation and POS tagging [C]// Proc of Conference on Empirical Methods in Natural Language Processing. 2013.

[7] Collobert R, Weston J. A unified architecture for natural language processing: deep neural networks with multitask learning [C]// Proc of the 25th International Conference on Machine Learning. Helsinki: International Machine Learning Society, 2008: 160-167.

[8] Chen Xinchi, Qiu Xipeng, Zhu Chenxi, *et al.* Long short-term memory neural networks for Chinese word segmentation [C]// Proc of Conference on Empirical Methods in Natural Language Processing. 2015: 1197-1206.

[9] Zhang Meishan, Zhang Yue, Fu Guohong. Transition-based neural word segmentation [C]// Proc of Meeting of the Association for Computational Linguistics. 2016: 421-431.

[10] Santos C N D, Xiang Bing, Zhou Bowen. Classifying relations by ranking with convolutional neural networks [J]. Computer Science, 2015, 86 (86): 132-137.

[11] Mikolov T, Chen Kai, Corrado G, *et al.* Efficient estimation of word representations in vector space [J]. Computer Science, 2013.

[12] Graves A. Long Short-term memory [J]. Neural Computation, 1997, 9 (8): 1735-1780.

[13] Zhao Hai. Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition [C]// Proc of the 6th SIGHAN Workshop on Chinese Language Processing. 2007.

[14] Cai Deng, Zhao Hai, Zhang Zhisong, *et al.* Fast and accurate neural word

segmentation for Chinese [C]// Proc of the 55th Annual Meeting of the Association for Computational Linguistics. 2017.

[15] Chen Xinchi, Shi Zhan, Qiu Xipeng, *et al.* Adversarial multi-criteria learning for chinese word segmentation [J]. Computer Science, 2017: 1193-1203.

[16] Huang Zhiheng, Xu Wei, Yu Kai. Bidirectional LSTM-CRF models for sequence tagging [J]. Computer Science, 2015.

[17] Lample G, Ballesteros M, Subramanian S, *et al.* Neural architectures for named entity recognition [C]// Proc of the North American Chapter of the Association for Computational Linguistics. 2016.

[18] Cai Deng, Zhao Hai. Neural word segmentation learning for Chinese [C]// Proc of the 54th Annual Meeting of the Association for Computational Linguistics. 2016: 409–420.

chinaXiv:201807.00065v1